Challenge: Image-Based Information Retrieval

Problem statement:

In today's digital world, organizations generate and process vast amounts of visual data, including scanned documents, images, and photos. However, extracting meaningful information from these images manually is time-consuming, prone to errors, and inefficient. This problem statement proposes the implementation of an automated information extraction solution from images, leveraging advanced technologies like Optical Character Recognition (OCR) and AI-based image analysis

Objective: Create a Python API that can take images as input, extract any embedded text, and return it in a clean and readable format. The API should be versatile enough to handle different image formats and various types of text content, including printed, handwritten, and stylized text.

Key Functionalities:

Image File Input:

- The API should accept common image formats, such as JPEG, PNG, TIFF, etc.
- Users should be able to upload images directly to the API for transcription.

Text Extraction:

- The core functionality of your API will be to extract text from the uploaded images.
- Implement Optical Character Recognition (OCR) to detect and transcribe text, regardless of the font style, size, or orientation.
- Handle various scenarios, such as multi-language text, complex layouts, and mixed content (images with both text and graphics).

Output and Formatting:

- The API should return the transcribed text in a clear and structured format.
- Provide options for different output formats, such as plain text, JSON, or CSV.
- Include metadata wherever applicable, such as the location of text within the image or the confidence level of the transcription.

Technical Expectations:

- **Accuracy**: Focus on delivering accurate transcriptions, especially for challenging text formats such as handwritten or low-quality scanned text.
- **Performance**: Ensure the API can handle images of various sizes and resolutions efficiently, providing results promptly.

- **Scalability**: Design the API to handle multiple image transcription requests simultaneously, with the ability to scale as needed.
- **Documentation**: Provide thorough documentation that explains how to use the API, including setup instructions, supported image formats, and example requests.

Suggested Tools and Technologies:

- Database: Postgresql
- Cloud Platforms: Azure
- Language: Python
- CI/CD Tools: Jenkins, GitLab CI/CD, GitHub Actions

Testing requirements:

- Functional Testing:
  To verify that all features of the information extraction solution work as intended.
  Scope: Test the core functionality, such as image upload, text extraction, data validation, and output generation.
- Accuracy Testing
  To measure the accuracy of the extracted information compared to the original content in the images.
- Performance Testing
  To assess the speed and efficiency of the information extraction process.
- Load Testing
  To evaluate the system's ability to handle large volumes of images and concurrent processing requests.
- End-to-End Testing
  To validate the complete workflow, from image input to final data output.
  Scope: Test the entire process, including image upload, preprocessing (if any), extraction, data validation, and output integration.
- Data Validation Testing
  To ensure the extracted data is accurate, complete, and formatted correctly according to the requirements.

Bonus Points:

- Support multi-language text detection and transcription.
- Implement features for recognizing and preserving text formatting, such as bold, italics, or underlined text.
- Include the ability to extract text from complex layouts, such as forms, tables, or images with overlapping text and graphics.

Submission Guidelines:

- Submit your solution including all necessary code, configuration files, and a README that explains how to set up and use your API.
- Provide examples showing the API's ability to transcribe text from various image types and formats.

Coding standards:

API coding standards:
https://digitaltechedge.sharepoint.com/:w:/r/sites/loopplatform/_layouts/15/Doc.aspx?sourcedoc=%7B5E288CE5-1F81-47CD-A912-7CA6F77A9A81%7D&file=Centific_Python_API_Guidelines_V1.0.docx&action=default&mobileredirect=true
Python coding standard:
https://digitaltechedge.sharepoint.com/:w:/r/sites/loopplatform/_layouts/15/Doc.aspx?sourcedoc=%7B71D98C50-4ED6-41E5-9CBD-EC751557C551%7D&file=Centific_Python_Coding_Standard_V1.0.docx&action=default&mobileredirect=true